様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成25年5月31日現在

機関番号：３２６３９
研究種目：若手研究（Ｂ）
研究期間：２００９～２０１２
課題番号：２１７００１７４

研究課題名（和文） 非合理な選択行動の特性から学習原理を導く

研究課題名（英文） Derivation of a learning principle from irrational behaviors

研究代表者
酒井 裕 （SAKAI YUTAKA）
玉川大学・脳科学研究所・准教授
研究者番号：70323376

研究成果の概要（和文）：ヒトを含め、動物は目先の利益にとらわれがちである。これは、動物が将来得られる利益の価値を主観的に割り引いており、より価値の高い利益を選択した結果であると解釈されている。このような主観的価値を取り入れた行動学習の枠組は強化学習理論として確立しているが、従来の枠組の適用範囲は狭く、動物行動に適用すると単純な実験例でも枠組が崩壊することを示した。そのため、この問題を解決する新たな行動学習の枠組を構築した。

研究成果の概要（英文）：Animals, including human, apt to prefer immediate returns. The learning framework incorpolating such a behavioral property is established as the reinforcement learing theory. However, we showed that the applicable range of the framework is small and outside of some simple examples of behavioral experiments of animals. We constructed a novel framework that is always applicable for animal behaviors and demonstrated that the proposed framework reproduces animal behaviors.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009年度	800,000	240,000	1,040,000
2010年度	700,000	210,000	910,000
2011年度	500,000	150,000	650,000
2012年度	500,000	150,000	650,000
総　計	2,500,000	750,000	3,250,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：遅延割引，学習行動，強化学習，割引価値問題

１．研究開始当初の背景

動物は自らの行動を選択し，その結果，生きるための報酬を得ている．得られる報酬を最大化しようとする際，脳にとっては知覚情報や自らの行動の履歴という膨大な情報を利用可能である．このような高次元の状態空間の中で標準的な学習アルゴリズムを用い，報酬最大化をしようとすると，膨大な試行回数が必要であることが知られている．しかし，実際の動物はいとも簡単に適切な行動選択を取れるように学習できる．膨大な情報の中から，報酬に関連するほんの一部の情報だけを抜き出し，その上で報酬最大化をしているとしか考えられない．これはどのような機構で実現しているのだろうか．

得られる報酬を最大化するために試行錯誤を通じて行動を最適化する枠組みは，強化学習理論として確立している[1,2]．しかし，強化学習理論では，環境に適した状態空間を明示的に与えられた上で，行動選択を最適化する枠組みしか定式化されておらず，環境に合っていない状態空間を設定した場合の理論は不十分であり，そもそもどのように適切な情報源を探索して，状態空間を設定するのか，という問題には一切触れられていない．

本課題の申請前に代表者らは，動物が報酬最大化に失敗する例として知られる「マッチング行動」[3]に注目し，この問題の解決の糸口にすることを試みた．

その結果，状態空間を実験課題に合わせて適切に設定できていないときに，ある性質を満たす多数の強化学習アルゴリズムが報酬最大化に失敗し，マッチング行動に至ることを示した[4,5]．この成果は状態空間の設定まで想定した場合に，従来の強化学習が見せるほころびを示していると共に，動物の行動学習の状況に合わせた新たな枠組みの必要性を示している．

２．研究の目的

本研究課題では，動物が報酬最大化に失敗する例，つまり客観的には非合理に見える動物行動に注目し，その行動特性から動物の脳で行われている学習原理を探り，動物がいかにして環境に適した情報源を選定し，状態空間を設定しているのか，という疑問に答える糸口を見つけることを目的とした．非合理行動として，申請以前から注目してきたマッチング行動に加え，遅延報酬に対する選好性[6]にも注目した．ヒトや動物は目先の利益にとらわれがちである．後で得られる大きな報酬より，目先の小さな報酬をしばしば選択する．一見，非合理に見えるこの選好性は，動物が主観的に将来の報酬の価値を割り引いており，主観的価値の高い報酬を選んだ結果である，と解釈されている[6]．強化学習理論の中では，平均獲得報酬を最大化する「平均報酬問題」と共に，将来報酬を割り引いた主観的価値を最大化する学習の枠組も「割引価値問題」として定式化されている[1]．本研究課題では，「平均報酬問題」と「割引価値問題」の両方の視点から，不適切な情報源にもとづいて設定した状態空間でも成立する学習の枠組みを構築し，マッチング行動や遅延報酬に対する選好性を再考し，その背後にある学習戦略を類推することを目的とした．具体的に掲げた目標は次の通りである．

(1) 任意の状態空間を想定した「平均報酬問題」の学習に関する理論的枠組を確立し，マッチング行動と遅延報酬に対する選好性を再考して，背後にある学習戦略を推定する．

(2) 任意の状態空間を想定した「割引価値問題」の学習に関する理論的枠組を確立し，マッチング行動と遅延報酬に対する選好性を再考して，背後にある学習戦略を推定する．

(3) 不適切な状態空間を設定してしまうような具体的な選択課題で，その振舞いから，「平均報酬問題」と「割引価値問題」のどちらのための学習戦略かを判別し，割引の程度を推定できるような選択課題を設計する．

３．研究の方法

本研究課題で目的とする枠組みと従来の強化学習の枠組みとの相違は，状態空間が予め与えられるのではなく，学習者自身で利用可能な情報源から設定する点にある．利用可能な情報源の中には，あらゆる感覚情報とその履歴が含まれ，これまでに自分自身が行った行動や得られた報酬も含まれる．膨大な情報であり，そのすべてからなる空間を状態空間とすると，同じ状態は通常２度と経験しないため，試行錯誤による学習は不可能である．したがって，膨大な情報源から，注目すべき情報のみからなる状態空間を設定していると考えられる．動物は必ずしも常に環境に適した状態空間を設定しているとは限らない，と想定し，行動学習の枠組みを再構築する．そのため，従来の強化学習理論の中の理論的基盤ひとつひとつを任意の状態空間でも成立するかどうか確かめ，成立する条件によって選り分けた．

行動選択は現在の状態に依存して行うため，ある状態空間を設定すると，あらゆる行動パターンの中で，限定された振舞いしか行えなくなる(図1)．したがって限定された範囲で行動を最適化することしかできない．あらゆる可能な行動パターンの中で最大の報酬を得られる真の最適行動を含むような状態空間は，次の条件を満たすような状態空間である[5]．

『環境に適した状態空間』
現在の状態を条件とした将来の報酬期待値が，過去に得られた情報に依存しない．

この条件は，利用可能な情報源の中から，将来得られる報酬に関する十分な情報を抽出して，現在の状態を捉えることができている，ということ意味する．動物は，この条件を満たすような状態空間の設定を目指しているが，あらゆる環境で実現するのは大変困難であると考えられるため，特に実験上，人工的な環境に置かれた場合，必ずしも適した状態空間を設定できていない，ということを想定する．

なお，ここで，各時刻に環境から得られる情報がその時点の状態だけである，という状況を想定すると，従来の強化学習が基盤としてきたマルコフ決定過程[1,2]と一致する．学習者自身が状態空間を設定する状況に合わせて，マルコフ決定過程を自然に拡張した条件になっている．

４．研究成果

(1)「平均報酬問題」の再構築

最適化の方法の中で，最も原始的で，しかも汎用性の高い方法として勾配法が挙げられる．試行錯誤を通じて勾配法を実現する学習アルゴリズムは「確率的方策勾配法」として知られている．平均的に勾配に比例するような更新則で，行動選択確率を更新するアルゴリズムである．勾配法を実現するために，平均報酬の勾配を計算すると，無限の積分が出てくる[7]．ある状態で，ある行動を行う確率に関する勾配は，その状態にいたときに，その行動を行ったか否かによって，将来の各時点で得られる報酬期待値がどの程度異なるか，という差分をあらゆる将来の時点に関して積分した値となる．

ある時点で行った行動が，将来どのくらいに渡って影響を及ぼすかわからないため，この勾配を正確に推定することは一般に困難である．この困難を回避するために，よく用いられる方法が時間差(Temporal Difference; TD) 学習である．TD学習では，予め各状態の価値を推定しておき，ある行動を行うことによって起きた状態遷移に伴う状態価値の差分を用いて，将来に渡る無限積分を置き換える．ある状態の状態価値は，その状態にいたとき将来の各時点で得られる報酬期待値が平均報酬に比べてどの程度か，という相対値をあらゆる将来の時点に関して積分したものと定義される．状態価値の推定にもTD学習が用いられ，状態遷移の際に変化した状態価値の差分によって，それ以降の無限積分を置き換える．

TD学習では，一見，行動や報酬に直接結びつかないような状態の価値も推定し，その価値をあたかも内的な報酬かのように用いる学習となっている．報酬との連合を学習しても決して得られる報酬が増えるわけではない古典的条件付けにおいて，なぜ報酬との連合が学習されるのか，という疑問に答える１つの解となっている．さらに古典的条件付けにおいて明らかにされている随伴性と状態価値が平均からの相対値であることは整合している．中脳黒質のドーパミン投射細胞の活動は，TD学習に用いられるTD予測誤差の振舞いと酷似していることから，TD学習が動物の脳の学習システムに実装されている可能性が示唆されている[8]．TD学習は，無限積分を回避するために，脳内で採用されている学習戦略として，有力な候補であると考えられる．

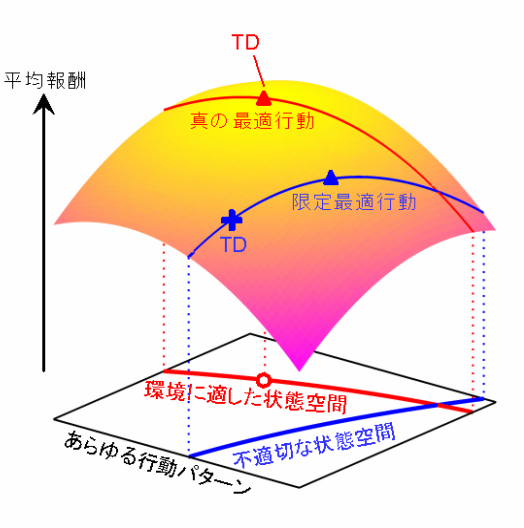

図1：状態空間の設定とTD学習

しかし，TD学習における状態価値の時間差分による置換は『環境に適した状態空間』を設定していなければ真とはならない．『環境に適した状態空間』を設定できておらず，限定された範囲で行動を最適化しようとしている場合，TD学習を用いると，その限定範囲の中での最適化もできなくなる(図1)．『環境に適した状態空間』を設定できていれば，TD学習による置換は真となり，真の最適行動に至ることができる．動物は『環境に適した状態空間』の設定を目指しており，その上でうまく働くTD学習を用いているのではないか．その結果『環境に適した状態空間』を設定できていない場合に，非合理に見える行動が顕れるのではないだろうか．

実際，マッチング行動も遅延報酬に対する選好性も『環境に適した状態空間』を設定できていない場合のTD学習の振舞いとして再現できることを示した．マッチング行動は，同一の状況での行動選択を繰り返し行い，過去の行動パターンに依存して確率的に報酬が与えられる課題で観測されている[3]．報酬確率が行動パターンにどのように依存しているのか，動物にはわかりにくい構造である場合が多い．その結果，もし行動選択をする状態が常に同一の状態となっているような状態空間を設定している場合，TD学習によってマッチング行動が顕れる．

動物の遅延報酬に対する選好性は，一定の量と遅延をそれぞれ割り当てた選択肢から一方を選ぶという試行を繰り返し行う選択課題で調べられている(図2)．

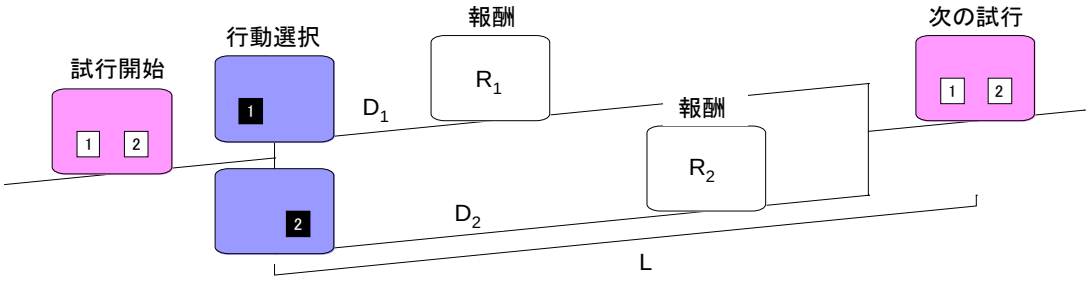

図2：遅延報酬選択課題

どちらを選んでも次の試行が始まるまでの時間を一定にしてあるため，遅延の長さによらず単純に報酬量の大きい選択肢を選ぶことが得である．しかし，動物は，各選択肢の遅延のパターンによって，小さい報酬の方を選ぶことがある．様々なパターンで動物の選好性を調べた結果，報酬量をR，遅延をDとすると，R/(1+kD)という値が大きい方を選ぶ，ということがわかっている．このことから動物が遅延報酬に対してR/(1+kD)という双曲型に割引された主観的価値を持っているという解釈がなされている[6]．しかし，このような選好性は，あくまで平均報酬を最大化しようとするTD学習の振舞いとしても再現される．この課題において，『環境に適した状態空間』となるためには，報酬を待っている遅延期間の間，今回どちらを選択し，どの程度時間が経っているか，という情報を反映した状態空間を設定しなければならない．しかし，行動選択をしてしまった後で，ただ待てばいい状態をこのように区別しておらず，遅延期間が同一の状態となるような状態空間を設定した場合，TD学習の振舞いとして，R/(1+kD)の大きい方を選択するようになることがわかった（〔雑誌論文〕③として発表)．

(2)「割引価値問題」の再構築

学習者自身が状態空間を設定することを想定し，将来報酬価値の割引を考慮した行動学習の枠組みを再考した．強化学習理論において，「割引価値問題」は割引状態価値を最大化する問題として定義されている．ある状態の割引状態価値とは，ある時点でその状態にいたとき，その後の各時点で得られる報酬に割引の程度をかけながら積分した値の期待値として定義される．割引状態価値は状態空間の各点で定義される値であり，複数存在する．この全ての割引状態価値を最大化する問題が「割引価値問題」である．

ある状態の割引状態価値は，その後訪れた別の状態でどのような行動選択を行うかに依存する．したがって，各状態ごとに独立には最大化できない．最大化すべき値が複数あることから，(1)のように通常の意味での勾配法を構成することができない．また，ある状態の割引状態価値を最大化しようとすると，別の状態の割引状態価値が最大でなくなる，というようなことが一般には起こりうるのではないだろうか．強化学習が基盤としているマルコフ決定過程では，このようなことが起こらず，全ての状態の割引状態価値を最大化するような行動の仕方が存在することが証明されており，その特性を活かした学習アルゴリズムが展開されている．しかし，本課題で取り組む任意の状態空間を想定した枠組みへと拡張していけるのだろうか．

この問題に取り組みながら，研究目的(3)を目指して，具体的な行動実験課題を設計してみた結果，動物実験でよく用いられる選択課題でも「割引価値問題」の最適解が存在しないような場合が起こることがわかった．すなわち，従来の強化学習理論で定式化されている「割引価値問題」は，任意の状態空間を想定した場合に拡張することができないことを示している．研究目的(2)(3)は，従来の「割引価値問題」に則って立てた計画であり，計画の変更が必要となった．また，従来の強化学習の「割引価値問題」に則った枠組みは分野を越えて神経科学，行動経済学，実験心理学に広まってきており，安易に適用範囲を越えて動物行動に当てはめることの問題点を指摘することは，分野への貢献として重要であると考えた．

(3) 研究計画の変更

研究目的(2)(3)を取りやめ，(4)従来の「割引価値問題」の枠組みを動物行動に適用することの問題点を指摘し，(5)新たな行動学習の枠組みを構築する．

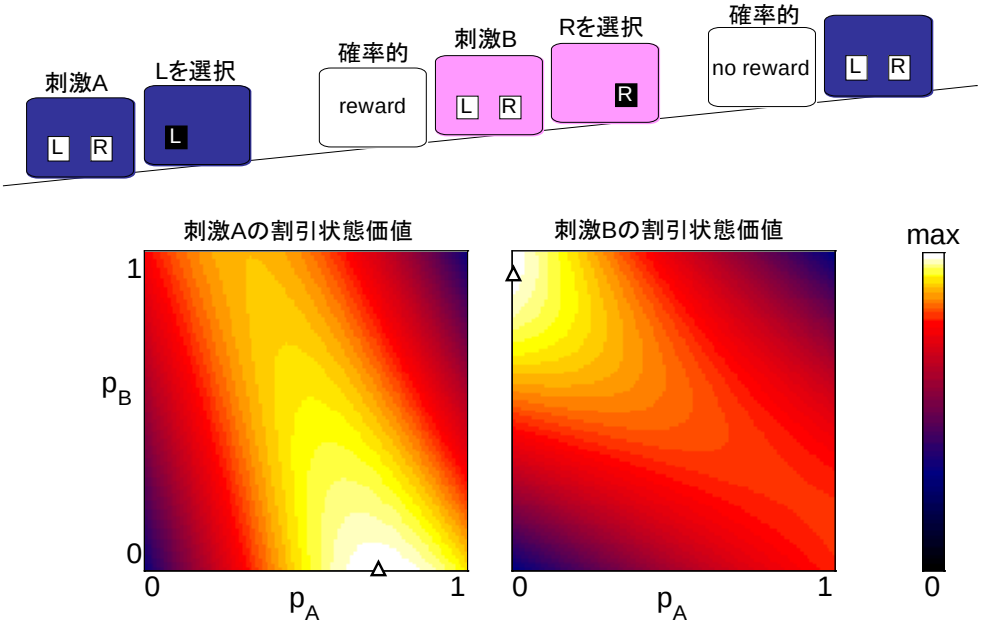

図3：「割引価値問題」の最適行動がない例

(4) 従来の「割引価値問題」の問題点

図3のように感覚刺激AもしくはBが提示され，左右にあるLかRのボタンを選ぶと，確率的に報酬が与えられる，という試行を繰り返す単純な選択課題を考える．このとき，まず動物が設定しそうな状態空間として，与えられる感覚刺激そのもの{A,B}が挙げられる．Aのときにボタンを選ぶ確率をp_Aとし，Bのときにボタンを選ぶ確率をp_Bとすると，それぞれRボタンを選ぶ確率は$1-p_A$, $1-p_B$で，(p_A, p_B)の組で行動選択の仕方が定まる．このとき，AとBの割引状態価値を両方とも最大化するような(p_A, p_B)の組にたどり着くことが，「割引価値問題」における目標となる．

報酬や感覚刺激の提示を決める確率ルールとして，あるルールを設定したときに，(p_A, p_B)の関数として，AとBの割引状態価値を求め，擬似カラーで図3に示した．Aの割引状態価値が最大になる(p_A, p_B)とBの割引状態価値が最大になる(p_A, p_B)が大きく異なっており，両方を最大化するような行動選択の仕方が存在しないことがわかる．ここで設定した課題の確率ルールでは，前回の試行に選んだボタンに依存して報酬確率も次に提示する感覚刺激も定めており，『環境に適した状態空間』は前回選んだボタン{L,R}となっている．感覚刺激{A,B}を状態空間として設定してしまうと，「割引価値問題」の解が存在せず，学習の目標を失ってしまうことがわかる．一体，何のために学習しているのかわからず，学習行動の理解の枠組みとして崩壊していることを意味する．

この欠陥が生じた原因を追究していくと，まず遅延報酬に対する動物の選好性を主観的価値として解釈する際，１試行分しか考慮していないことに問題があると考えられる．強化学習で次々と状態が遷移していく状況に適用するために，それぞれの状態で主観的価値を定義する必要が生じた結果，最大化すべき値が複数になってしまった．また，試行をまたいだ先まで考慮すると，別の欠陥も生じる．将来報酬が複数の時点で得られる場合，全体の主観的価値はそれぞれの和となるのが妥当である．しかし，双曲割引型の主観的価値 $R/(1+kD)$は，遅延Dに関して積分すると発散するため，将来の報酬確率が0とならない限り，主観的価値の総和は発散する．現実的な遅延時間までで総和を制限したとしても，この発散特性から，主観的価値に対する重みは長期的な将来報酬の方が大きくなり，目先の小さな報酬を選択するような行動は顕れないはずである．

(5) 新たな「割引価値問題」の構築

これらの欠陥を解決する糸口として，ヒトの繰り返し試行の実験で，試行をまたいだ割引の効果は双曲型とならず，指数型 $R\gamma^D$ となるという知見[9]に注目した．遅延期間に待っている間の割引の効果と試行をまたぐときの割引の効果は，異なる可能性があり，この両者をつなぐためには，連続的な時間と離散的な時間ステップの両方の特性を導入する必要がある．そこで，イベント発生によって進む離散的な時間ステップを導入し，そのイベント間の間隔は変動することを想定した．これはセミマルコフ決定過程[1,2]で導入されている時間ステップと同等である．

ここでは，従来の「割引価値問題」の欠陥を解決するため，最大化すべき値は１つであるような割引価値で，動物の選好性を再現するものは何か，検討した．その結果，イベントの間は連続的な実時間に従って双曲型の割引が起こり，イベントをまたいで時間ステップが進むときには，掛け算で割引の効果が起こるような主観的価値を定義すると，動物の選好性を再現することがわかった．さらに，イベントが等間隔である場合には，時間ステップにしたがって，指数型の割引となり，知見[9]とも整合する．最大化すべき値が１つであるため，(1)「平均報酬問題」と同様に勾配法を導出することができ，同様にTD学習による最適化が可能である（(4)(5)は〔雑誌論文〕①として発表）．

(6) 位置づけと今後の展望
当初，従来の強化学習理論で定式化されている「割引価値問題」の枠組みを元にして，学習者自身が状態空間を設定するということを想定した理論的な拡張と具体的な選択課題の設計を目指し，研究計画を立てたが，従来の「割引価値問題」はそのまま拡張することができないことが判明し，計画の変更を余儀なくされた．しかし，その「割引価値問題」の欠陥を指摘したことは，分野を越えてインパクトを与えると考えられる．強化学習理論の中で，「平均報酬問題」より「割引価値問題」に則った枠組みの方が広まっており，分野を越えて神経科学，行動経済学，実験心理学で実験データのモデルベース解析に利用されるようになってきている．しかし，そのほとんどが，適用条件を満たさない動物行動に当てはめており，その問題点が認識されていない．そのような状況に警鐘を鳴らし，適切な枠組みを提供する研究成果である．
　今後は当初の目的どおり，ヒトや動物が環境に適した状態空間を設定できない場合の行動を調べていく．ラットの頭を固定した統制環境に置き，行動実験する準備を進めている（〔雑誌論文〕②として発表）．今後はヒトや動物が如何にして状態空間の設定を行っているのか探求していく．

[1] Bertsekas & Tsitsiklis, Neuro-Dynamic Programming, 1996
[2] Sutton & Barto, Reinforcement Learning, 1998
[3] Herrnstein, The Matching Law, 1997
[4] Sakai et al. Neural Netw, 2006
[5] Sakai & Fukai, PLoS One, 2008
[6] Mazur & Biondi, J Exp Anal Behav, 2009
[7] Williams, Machine learning, 1992
[8] Schultz et al. Science, 1997
[9] Schweighofer et al. PLoS Comput Biol, 2006

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔雑誌論文〕（計3件）
① Yamaguchi Y, Sakai Y, Reinforcement learning for discounted values often loses the goal in the application to animal learning, Neural Networks, vol.35, pp.88-91, 2012, 査読有, DOI: 10.1016/j.neunet.2012.08.004

② Kimura R, et al. Sakai Y (8人中7番目), Reinforcing operandum: rapid and reliable learning of skilled forelimb movements by head-fixed rodents, Journal of Neurophysiology, vol.108, pp.1781-1792, 2012, 査読有, DOI: 10.1152/jn.00356.2012

③ Yamaguchi Y, Sakai Y, A theoretical approach to animal's impulsive preference: Impulsive choice behavior is interpreteded as a result of reward-maximization failure, Proceedings of SCIS-ISIS2012, Kobe, Japan, Nov. 20-24, pp.1182-1185, 2012, 査読有, DOI: 10.1109/SCIS-ISIS.2012.6505271

〔学会発表〕（計7件）
① Yamaguchi Y, Sakai Y, Impulsive preference emerges as a result from reward-maximization failure, 第35回日本神経科学大会, 2012/9/19, 名古屋
② 山口良哉, 酒井裕, 割引価値問題は被験者の戦略によって不良設定問題となる, 日本神経回路学会 第21回全国大会, 2011/12/15, 沖縄
③ 酒井裕, 非合理行動の背後にある合理的学習戦略, 日本基礎心理学会 第30回大会, 2011/12/4, 横浜
④ Yamaguchi Y, Sakai Y, Discounted value problem becomes ill-posed by subject's strategy, 8th IBRO, July 16, 2011, Florence, Italy,
⑤ 山口良哉, 酒井裕, 強化学習における時間割引の再考, 日本物理学会 第66回年次大会, 2011/3/25, 新潟
⑥ Yamaguchi Y, Sakai Y, Purpose or Strategy? Reconsideration of temporal discount in non-Markov situation, 第33回日本神経科学大会, 2010/9/4, 神戸
⑦ Yutaka Sakai, A hypothesis of efficient learning rule: dopamine-dependent metaplasticity, 第33回日本神経科学大会, 2010/9/3, 神戸

〔その他〕
ホームページ等
http://spike.lab.tamagawa.ac.jp/

６．研究組織
(1)研究代表者
酒井 裕（SAKAI YUTAKA）
玉川大学・脳科学研究所・准教授
研究者番号： 70323376